UNIVERSITÉ DE MONTRÉAL

CENTRE DE RECHERCHE SUR LES TRANSPORTS

CASE POSTALE 6128, SUCCURSALE A, MONTRÉAL, QUÉBEC, H3C 3J7

ON OPTIMAL ROUTE CHOICE STRATEGIES IN TRANSIT NETWORKS

by

Heinz Spiess

Ce rapport a été publié grâce à une subvention du fonds F.C.A.C. pour l'aide et le soutien à la recherche.

Université de Montréal

Centre de recherche sur les transports - Publication #286
Département d'informatique et de
recherche opérationnelle - Publication #466

Février 1983

Dépôt légal - 1er trimestre 1983

Abstract

We describe a model for the transit assignment problem with a fixed set of transit lines. The traveler chooses the strategy that allows him to reach his destination at minimum expected cost. First we consider the case in which no congestion effects occur. If no assumptions are made on the distribution of waiting times, the model may be stated as a linear programming problem of a size that increases exponentially with the size of the transit network. For the special case in which the waiting time at a stop depends only on the combined frequency, the problem is formulated as a linear programming problem of a size that increases linearly with the network size. An algorithm is developed that solves the latter problem in polynomial time. The model is then extended to the case in which the travel time on a link is an increasing function of the link volume. Equilibrium conditions for route choice are stated and it is shown that the problem is equivalent to a convex mininization problem which may be solved by an adaptation of the Frank and Wolfe method.

Résumé

Nous décrivons un modèle pour l'affectation du transport en commun sur un réseau de lignes à itinéraires fixes. Le voyageur choisit la stratégie qui lui permet d'atteindre sa destination à coût minimum. D'abord nous considérons le cas où les effets de congestion sont négligeables. Si on ne fait pas d'hypothèses sur la distribution du temps d'attente, le modèle peut être formulé comme un problème de programmation linéaire dont les dimensions augmentent exponentiellement avec les dimensions du réseau considéré. Dans le cas spécial où le temps d'attente ne dépend que de la fréquence combinée, on peut reformuler le problème comme un programme linéaire dont les dimensions sont proportionnelles à celles du réseau de transport en commun. Nous développons un algorithme qui peut résoudre ce dernier problème dans un temps polynominal. Le modèle est ensuite généralisé au cas où le temps de voyage sur un lien est une fonction croissante du volume sur ce lien. Nous donnons des conditions d'équilibre et nous montrons que ce problème est équivalent à un problème de minimisation convexe qui peut se résoudre par une adaptation de la méthode de Frank et Wolfe.

1. Introduction

The purpose of this paper is to describe a new model and solution algorithm for the transit assignment problem on a network that consists of fixed transit lines. In the first part of the paper the case in which congestion effects do not influence the traveler's route choice is considered. While the corresponding problem for car traffic is solvable by an "all-or-nothing" assignment to the shortest paths between origins and destinations, the problem becomes much more complex in the case of a transit network, due to the phenomenon of waiting for the arrival of a vehicle which becomes an element of the traveler's route choice decision. Also, in the case of the car traffic the problem is symmetric with respect to origin and destination. This symmetry does not hold in the transit assignment problem.

The transit assignment problem has been studied by several authors in the past, either as a separate problem (e.g. see Dial [6], Chriqui [4], Rapp et al. [19], Andreasson [1]) or as a subproblem of more complex models, such as transit network design (see e.g Lampkin and Saalmans [14], Scheele [20], Mandl [17] and recently Hasselström [13]), or multimodal network equilibrium (Florian [9], Florian and Spiess [10]). Computer programs, such as IBTDS [12], TRANSCOM [3], TRANSEPT [15], NOPTS [19], the VOLVO system [1,13] and ULOAD [21], are each based on one of these methods and are widely used by transit companies and government agencies

for planning purposes.

All the algorithms that have been proposed in the past may be classified as heuristic approaches to this problem and are not based on a behavioral principle that resulted in the formulation of a model, rather the algorithms are stated directly and are in most cases variants of assignment techniques normally used for car traffic on road networks (e.g. shortest path, multipath assignment), that are adapted to reflect some of the phenomena inherent to transit networks. Although for some of these algorithms the corresponding problem that is solved may be identified, these are restricted versions of the general problem presented here, and therefore only yield suboptimal solutions.

The subsequent sections of the paper are organized in the following way. In Section 2 we discuss a model of the transit traveler's behavior which leads to the concept of the "optimal strategy". This allows us to put in perspective and compare the existing methods and point out the restrictions that are embedded in each of these algorithms. In Section 3 the mathematical programming model, that corresponds to the postulated traveler's behavior according to an optimal strategy, is developed. We show that the problem may be stated as a linear program of a size that increases exponentially with the size of the transit network. In order to keep the notation as simple and general as possible, the problem is stated for a network in which

every link is subject to waiting. In Section 4 we examine the special case in which the waiting time at a boarding point depends only on the combined frequencies of the links considered at this node. We show that the problem may then be restated as a simpler linear program of a size that increases only linearly with the network size. An algorithm to solve this special version of the problem is presented in Section 5. It is shown that the algorithm has polynomial complexity. In Section 6 the model is extended to the case in which the link times are functions of the link volumes. This allows us to model some of the aspects of vehicle capacity and congestion. Equilibrium conditions, an equivalent convex optimization problem and a solution algorithm are discussed. Finally, Section 7 contains some conclusions that may be drawn from this work and outlines its potential for future research in related areas.

2. Strategies in Transit Route Choice

Let T denote a transit network consisting of a set of nodes, a set of transit lines, each defined as a sequence of nodes at which passengers may board and alight, and a set of walk links, each connecting two nodes. The times (or costs) that are associated with each walk link and each segment of a transit line are constant and known. At each node that is served by a line, the distribution of interarrival times of the vehicles of that line is known. Therefore, assuming that we also know the rate of arrival of the passengers at each node, we can construct the distribution of the waiting time for a vehicle of a given line. Hence, we can calculate the combined expected waiting time for the arrival of the first vehicle, among any given set of lines passing at the same node, as well as the probability of each line arriving first. The details of these calculations are given in Appendix A.

Note that each walk link may be replaced by a transit line of one link with a zero waiting time (very high frequency). In the following, the walk links will therefore not always be distinguished from the segments of the transit lines. In order to simplify the formulation, it is also assumed that the transit network T is strongly connected.

We now turn our attention to the transit traveler whose trip originates at node A and is destined to node B. We assume that he behaves in a way that minimizes his total

expected travel time, which is the sum of expected walk, wait, and in-vehicle time. This assumption seems natural and it also parallels the behavioral assumptions made for car traffic route choice. It is possible, and probably very realistic, to assign different weights to the different components of the travel time. Since this may always be achieved by scaling the times used in the definition of the network, we need not consider this case separately. Thus, we will always use the term time, even if these 'times' might actually include weight factors and should then more appropriately be called 'generalized costs'.

One could be tempted to reformulate the above problem as: "How does one find the path from node A to node B which minimizes the expected total travel time?", and by doing so one would already have changed the problem. We are accustomed to associate the choice set of a traveler with the set of all possible paths. This is, of course, true in the case of a car driver, but this is not the case for a transit rider. He might, for instance, choose a set of paths and let the vehicle that arrives first decide which of the paths is actually used to get to his destination; in which case his choice does not correspond with the specification of a path. His choice set is therefore bigger than just the set of possible paths. If the problem is limited to finding the path with the smallest expected travel time, a restriction of the problem is solved, yielding a suboptimal solution.

We call any element of the traveler's choice set a *strategy*. A strategy is a set of rules that, when applied, allows the traveler to reach his destination. The number and the kind of different strategies that the traveler may choose from depend on the information that is available to him during his trip. If no additional information becomes available during the trip, a strategy simply defines a path. For the transit network given in Figure 1, an example of such a strategy would be:

> "Take line 2 to node Y; transfer to line 3 and exit at node B."

If the traveler, while waiting at a node, knows the line which is to be served next, a strategy may include subsets of lines and would then be:

> "Take the next vehicle among lines 1 and 2; if line 1 was taken, exit at B; if line 2 was taken, transfer at node Y to any of the lines 3 or 4 and exit at B"

Of course, the more information available to the traveler (such as elapsed waiting time, arrival times of vehicles seen while waiting at a stop, information on other vehicles while riding in a vehicle by looking out of the window, ...), the more complex the strategies may become; for example:

> "Wait up to five minutes for a vehicle of line 1; otherwise take line 2; if at node Y you see a vehicle of line 3 (express bus), transfer to line

3, otherwise continue to node X and transfer there to any of the lines 3 or 4."

We make the assumption that the only information that is available to the traveler during his trip is that he knows, while waiting at a node, which line is served next. Consequently, we only consider strategies that correspond to the second example given above. It is then possible to define a strategy of this type by specifying at every node of the network, except at the destination node, a non-empty set of attractive lines, and for each of these lines the (upstream) node at which the traveler alights. In the remainder of this paper we only consider strategies of this type. Note that when walk links are included in the network, a strategy may contain at a node an outgoing walk link instead of the set of attractive lines. Given a strategy, an actual trip is then carried out according to a mechanism that may be stated systematically as:

0: Set NODE to origin node.

1: Board vehicle that arrives first among the vehicles of the set of attractive lines at NODE.

2: Alight at the predetermined node

3: If not yet at destination, set NODE to current node and return to step 1. Otherwise trip is completed.

A strategy is said to be *feasible* if the graph defined by the attractive lines of the strategy does not contain cycles. Note that a strategy as defined above has a well determined destination node. The origin node, on the other hand, does not enter in the definition of a strategy. A strategy is therefore a rule that enables the traveler to reach his destination starting from any node in the network.

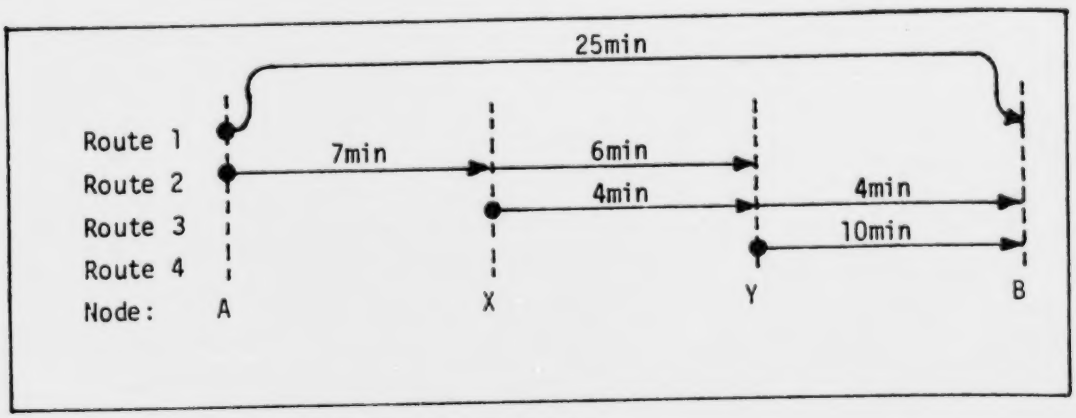

Figure 1: An example transit network.

Figure 1, which represents the transit lines used earlier to illustrate the concept of a strategy, shows a small transit network consisting only of four nodes and four transit lines. Since the network does not contain cycles, any strategy, as defined above, is feasible. There are 75 different strategies for reaching the destination node B. This is seen by the enumeration of all possible sets of attractive lines at each node except B, as showed in Table 1. Since any combination yields a feasible strategy, there are $5 \times 5 \times 3 = 75$ distinct feasible strategies to reach node B. However, there are only five different paths from A to B.

Node	Attractive Routes [Route -> Exit Node]	Waiting Time [Minutes]	Route Probabilities R1	R2	R3	R4
A	1->B	6.0	1.00	-	-	-
A	2->X	6.0	-	1.00	-	-
A	2->Y	6.0	-	1.00	-	-
A	1->B, 2->X	3.0	0.50	0.50	-	-
A	1->B, 2->Y	3.0	0.50	0.50	-	-
X	2->Y	6.0	-	1.00	-	-
X	3->Y	15.0	-	-	1.00	-
X	3->B	15.0	-	-	1.00	-
X	2->Y, 3->Y	4.3	-	0.71	0.29	-
X	2->Y, 3->B	4.3	-	0.71	0.29	-
Y	3->B	15.0	-	-	1.00	-
Y	4->B	3.0	-	-	-	1.00
Y	3->B, 4->B	2.5	-	-	0.17	0.83

TABLE 1: All Possible Sets of Attractive Routes

By using the in-vehicle times indicated in Figure 1 and the expected waiting time and line probabilities for each set of attractive lines at each node, as given in Table 1, we can now calculate the expected total travel time from A to B for each strategy. Figure 2 shows the optimal strategy, that is, the one minimizing the expected travel time. Note that the expected travel time is 27.75 minutes and that this strategy does not correspond to the choice of a single path.

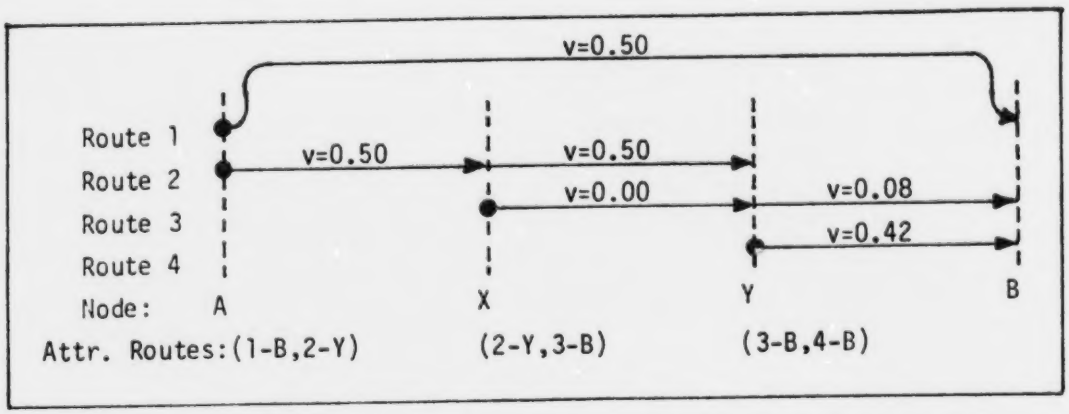

Figure 2: Optimal strategy. Expected travel time A to B is 27.75 min.

If we compare this strategy with the one found by algorithms that are based on finding the least cost path (Dial [6], Fearnside and Draper [8]), we see that the solution found by these methods, given in Figure 3, is clearly inferior to the optimal solution. The method described by Rapp et al. [19] is also based on finding the shortest path, but the trips are then assigned to all efficient (in the case of the example all) paths according to Dial's multipath assignment method [7]. Since the waiting times are those of individual lines (i.e. combined waiting times are not used), this method yields a solution with an expected travel time that is worse than the one of the least time path.

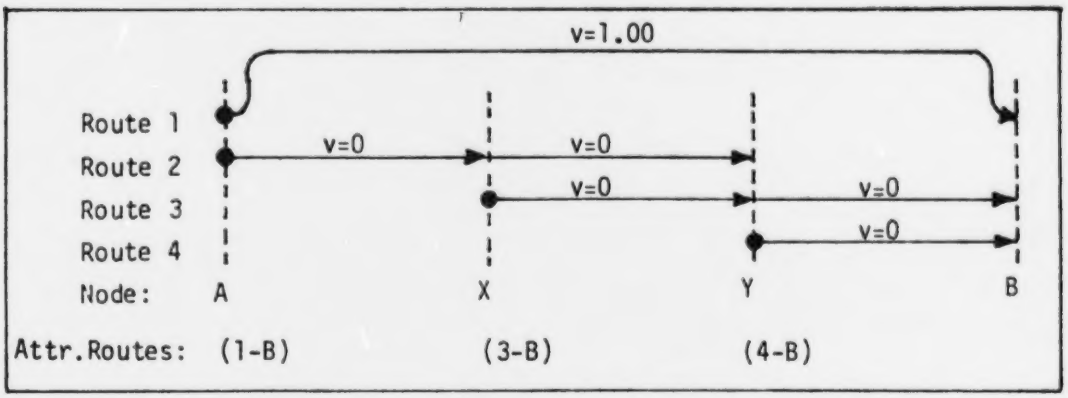

Figure 3: Best path to node B. Expected travel time A to B is 31 min.

Other methods described in the literature (Le Clercq [16], Chriqui [4]) are also based on the idea of a finding a least time path. But instead of defining the path in terms of consecutive segments of individual lines, paths are defined as a sequence of transfer nodes. Between two transfer nodes, the traveler's strategy may include more than one of the lines that connect the two transfer nodes directly. While in Le Clercq's method all direct lines are considered to be attractive, regardless of their in-vehicle time, Chriqui solves the problem of selecting the optimal subset of direct lines analytically (see Chriqui and Robillard [5]). The solution obtained by Chriqui's algorithm, which coincides for this example with Le Clercq's solution, is displayed in Figure 4. However, both of these models do not handle strategies of the general type.

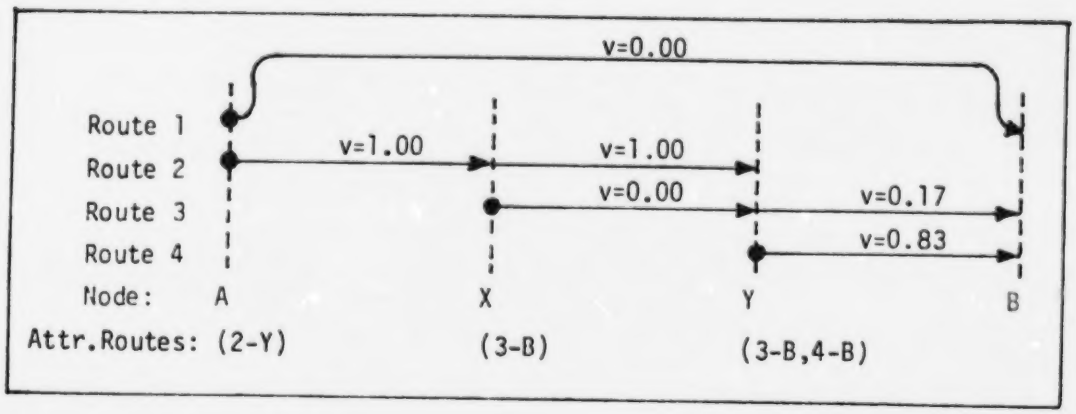

Figure 4: Best transfer sequence. Expected travel time A to B is 30.5min.

Finally, the method proposed by Andreasson [1] and later refined by Hasselström [13], enumerates a certain number of distinct paths. Among these, the attractive paths are selected by a rather complicated set of heuristic rules that seem to work quite well in practice but do not guarantee optimality of the expected travel time. While all other methods mentioned above assign the demand from one origin to all destinations in one step, this method assigns the demand of each origin - destination pair separately.

In the sections that follow, a model is developed in which the traveler's choice set consists of all feasible strategies. It will become evident that, due to the asymmetry inherent in the transit trip, this model can only be solved efficiently by assigning the demands from all origins to a single destination at the same time.

3. The Model

As we have seen in the preceding section, a transit trip consists in general of several trip components that may include some or all of the following:

- access from origin to transit stop,
- waiting for a vehicle,
- riding in a vehicle,
- alighting a vehicle,
- walking between two transit stops and
- egress from transit stop to destination.

A non-negative time (or cost) is usually used to quantify each of these trip components, with the exception of "waiting for a vehicle", which is quantified by using the statistical distribution of waiting times for the arrival of the first vehicle of a given transit line at a given stop.

In order to simplify the presentation, we consider in the following a generalized formulation of the problem, in which every trip component includes a constant time as well as a distribution of waiting times. The trip components are represented by links $l \in L$ of a network $G=(I,L)$ with nodes $i \in I$. Let L_i^+ (L_i^-) denote the set of outgoing (incoming) links at node $i \in I$. Each link $l \in L$ is characterized by the pair (t_l, Θ_l) where t_l is a non-negative link travel time and Θ_l the distribution function for the waiting time

$G_l(x)$ = Prob [waiting time on link $l \leq x$].

The functions $G_l(x)$ can be obtained from the distribution of interarrival times (headways) and the distribution of passenger arrival times (see Appendix A for details). For a link l that does not involve waiting we have

$$G_l(x) = D(x) = \begin{cases} 0, & \text{if } x < 0, \\ 1, & \text{if } x \geq 0. \end{cases}$$

Using this generalized formulation we can represent any transit network T. The mechanics that are involved in constructing the links and the nodes of the generalized network depend very much on the particularities of the transit network and the degree of aggregation considered.

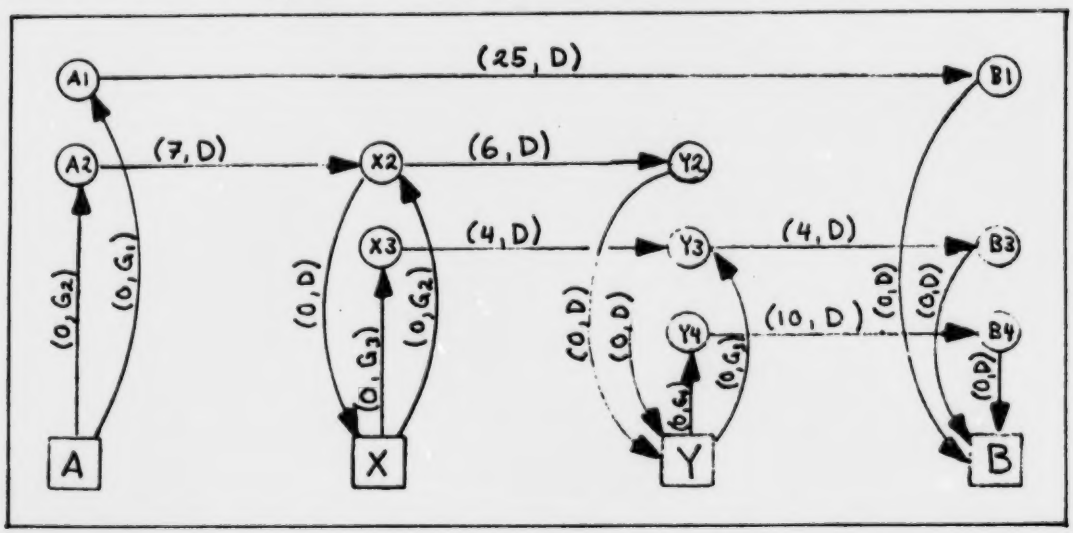

Figure 5: Link representation of example transit network.

Figure 5 shows such a network representation of the transit network used as an example in section 2. In this case only three types of trip components are considered: waiting (no travel time), riding (no waiting) and alighting

(no travel time and no waiting time). By noting that a node which has only one incoming link l_1 with (t_{11},θ_{11}) and one outgoing link l_2 with $(t_{12},\theta_{12}=D)$ may be replaced with a single link l with $(t_l=t_{11}+t_{12},\theta_l=\theta_{11})$, the network of Figure 5 may be reduced to the one shown in Figure 6.

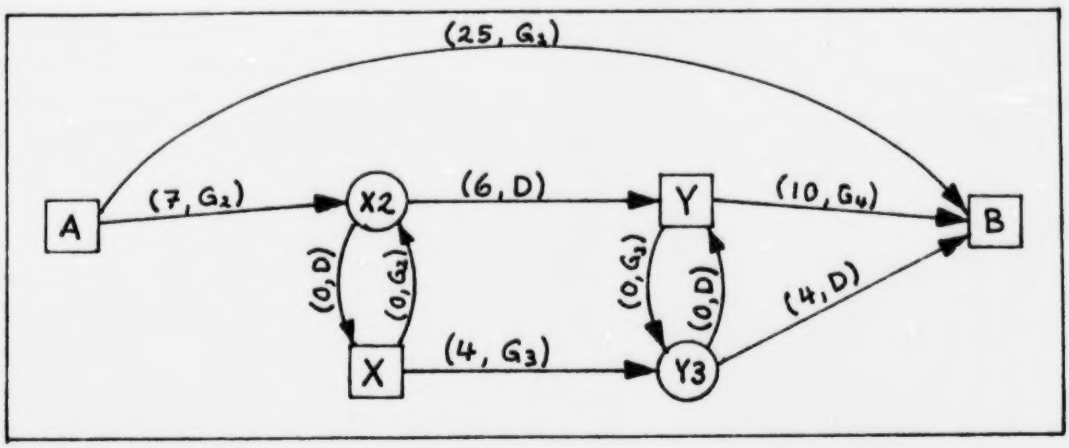

Figure 6: Simplified link representation of example transit network.

Since this generalized formulation no longer explicitly identifies transit lines but relies only on links, we can no longer use the terminology associated with the transit lines, such as waiting for a vehicle of a transit line. Instead we will use in the following the convention that each link is *served* by vehicles and that the passenger, instead of waiting for a vehicle of a certain line, waits for a *link to be served*.

For the transit route choice problem in this generalized form, a strategy to reach destination node r is defined by a partial network $\theta_r=(I,\bar{L})$ that contains only those links

that will be used as a consequence of this strategy. We can therefore denote a strategy by the corresponding subset of links $\bar{L} \subseteq L$, or alternatively, by the partition of these links according to their i-node ($\bar{L}_i^+ = L_i^+ \cap \bar{L}$, $i \in I$). Links not included in strategy $\bar{L}$ are never used. Among the links that are included in the strategy $\bar{L}$, at each node $i \in I$, a traveler boards the first vehicle that serves any of the links $l \in \bar{L}_i^+$. Hence, $\bar{L}_i^+$ corresponds to the set of attractive lines that was discussed in section 2 and, of course $\bar{L}_i^+ \neq \emptyset$ for $i \neq r$. For a strategy to be *feasible* it must contain at least one path from each node to the destination node r and it must not contain any cycles.

Let $W(\bar{L}_i^+)$ denote the expected waiting time for the arrival of the first vehicle serving any of the links $l \in \bar{L}_i^+$. $W(\bar{L}_i^+)$ is called the *combined waiting time* of links $l \in \bar{L}_i^+$. Let further $P_l(\bar{L}_i^+)$ be the probability that link l is served first among the links $\bar{L}_i^+$. For convenience we define $P_l(\bar{L}_i^+) = 0$ for all $l \notin \bar{L}_i^+$. Of course $W(\bar{L}_i^+)$ and $P_l(\bar{L}_i^+)$ depend on the distributions of waiting time $B_l(x)$. The exact formulae are given in Appendix A.

Let g_i, $i \in I-\{r\}$, be the demand (number of trips) from node i to the destination node r. In order to simplify the notation, we define

$$g_r = - \sum_{i \neq r} g_i .$$

We assume further that $g_i > 0$, $i \in I$, which prevents

degeneracy of the variables to occur, thus allowing for more compact proofs in the analysis of the model. We note however, that all results presented below remain valid for $g_i \geq 0$, $i \in I$.

Given a strategy $\bar{L}$, the demand g_i, $i \in I$, is assigned to the network yielding link volumes v_l, $l \in L$, and node volumes v_i, $i \in I$. The volume at a node is the sum of the volumes of all incoming links and the demand at that node

$$v_i = \sum_{l \in L_i^-} v_l + g_i \qquad , i \in I. \tag{1}$$

The node volume v_i is distributed on the outgoing links according to their link probabilities under strategy $\bar{L}$

$$v_l = P_l(\bar{L}_i^+)\, v_i \qquad , l \in L_i^+,\ i \in I. \tag{2}$$

Since $\sum_{l \in L_i^+} P_l(\bar{L}_i^+) = 1$, (1) and (2) ensure that the conservation of flow equation

$$\sum_{l \in L_i^+} v_l - \sum_{l \in L_i^-} v_l = g_i \qquad , i \in I \tag{3}$$

is satisfied.

The optimal strategy $\bar{L}^*$ is the strategy which minimizes the expected total travel time including waiting time. It is found by solving problem P1 for v_i^*, $i \in I$, v_l^*, $l \in L$, and $\bar{L}^*$:

$$\text{Min} \quad \sum_{l \in L} t_l\, v_l + \sum_{i \in I} W(\bar{L}_i^+)\, v_i \tag{P1}$$

subject to (1), (2) and

$$v_i \geq 0 \qquad ,i \in I. \tag{4}$$

The non-negativity of the link volumes v_l need not be stated explicitly, since it follows directly from (1) and (4).

Problem P1 is stated in an unusual way, in the sense that the set of decision variables includes not only the continuous variables v_i and v_l but also the link subset $\bar{L}$. It is possible to reformulate problem P1 as a linear programming problem by eliminating the explicit choice of $\bar{L}$. For that purpose, let K_i, $i \in I$, be the set of all possible non-empty link subsets $k = \bar{L}_i^+ \subseteq L_i^+$. Let w^k and p_l^k, $l \in L_i^+$, denote the expected combined waiting time and the link probabilities for the subset $k \in K_i$. The node volumes v_i are split into components v^k, corresponding to all possible link subsets $k \in K_i$. We can now formulate the linear programming problem P2

$$\text{Min} \quad \sum_{l \in L} t_l v_l + \sum_{i \in I} \sum_{k \in K_i} w^k v^k \tag{P2}$$

subject to

$$v_l = \sum_{k \in K_i} p_l^k v^k \qquad ,l \in L_i^+, i \in I, \tag{5}$$

$$\sum_{k \in K_i} v^k = \sum_{l \in L_i^-} v_l + g_i \qquad ,i \in I, \tag{6}$$

$$v^k \geq 0 \qquad ,k \in K_i, i \in I. \tag{7}$$

Proposition 1: Problems P1 and P2 are equivalent.

Proof: First note that every feasible solution $(v,\bar{L})$ of problem P1 corresponds to an extreme point of problem P2, with the same value for the objective function, of the form

$$v'_l = v_l \qquad ,l \in L,$$

$$v^k = \begin{cases} v_i & \text{if } k=\bar{L}_i^+ \\ \\ 0 & \text{else} \end{cases} \qquad ,k \in K_i, i \in I.$$

On the other hand, every extreme point of problem P2 has at most as many k with $v^k>0$ as there are nodes, and that (6), together with the positivity of the demands g_i, implies that for each node $i \in I$, there is at least one $k \in K_i$ with $v^k>0$. Thus, every extreme point of P2 corresponds to a strategy $\bar{L}$ and therefore to a feasible solution of P1. Hence, the feasible set of P2 is the convex hull of the lattice points associated with the feasible solutions of P1. Since any linear program is equivalent to its restriction to extreme points, P2 is equivalent to P1.□

Having transformed the original problem P1 into a linear programming problem, we can examine the corresponding dual problem P3, which is

$$\text{Max} \quad \sum_{i \in I} g_i u_i \qquad \text{(P3)}$$

subject to

$$u_j + t_l + m_l = u_i \qquad ,l=(i,j)\in L, \tag{8}$$

$$\sum_{l\in L_i^+} p_l^k m_l \leq w^k \qquad ,k\in K_i, i\in I. \tag{9}$$

The dual variables u_i are the expected total travel times to reach destination r from node i, as can be seen from the objective function, which, at its optimum, assumes the same value as the objective function of the primal problem P2. Since P3 is invariant to translations of the form $u_i'=u_i+c$,$i\in I$, we can assume $u_r=0$ without loss of generality.

Proposition 2: Any optimal solution (u^*,m^*) of problem P3 corresponds to a strategy $\bar{L}^*$ which satisfies

$$\sum_{l\in L_i^+} P_l(\bar{L}_i^{+*}) m_l^* = W(\bar{L}_i^{+*}) \qquad ,i\in I. \tag{10}$$

Proof: Any $\bar{L}^*$ obtained as optimal solution of problem P1 satisfies (10), as is easily seen by applying proposition 1 and the theorem of complementary slackness of linear programming. □

The variables m_l have no immediate interpretation. However, using (10) m_l may be interpreted as waiting time components, distributed among the links $l\in\bar{L}_i^{+}$ that yield a weighted total corresponding to the combined waiting time $W(\bar{L}_i^{+})$. The upper bound for the variables m_l

$$m_l \leq W(1) \qquad ,l\in L \tag{11}$$

follows directly from (9), since every link l is itself a

subset of L_i^+ and corresponds therefore to a $k \epsilon K_i$.

Note that the variables m_1 must not be confused with the conditional expected waiting time for the first vehicle serving any of the links $\bar{L}_i^+$, subject to link 1 being served first.

4. Special Case of Combining Frequencies

In the following, we consider the special case in which the waiting time distribution of each link l (or transit line, in the original form of the problem) is quantified by a positive parameter f_l, which will be called the *frequency* of a link. The expected combined waiting time and the link probabilities are derived from the frequencies in the following way

$$W(\bar{L}_i^+) = \frac{a}{\sum_{l \in \bar{L}_i^+} f_l} \qquad ,a>0, \tag{12}$$

$$P_l(\bar{L}_i^+) = \frac{f_l}{\sum_{l' \in \bar{L}_i^+} f_{l'}} \qquad ,l \in \bar{L}_i^+. \tag{13}$$

The case a=1 corresponds to an exponential distribution of interarrival times of the vehicles with mean $1/f_l$ and a uniform passenger arrival rate at the nodes.

The case a=1/2 is an approximation of a constant interarrival time $1/f_l$ for the vehicles on link l. This measure of waiting time is the most widely used approach in practice [3,6,16,19,21], in spite of the fact that it is based on a rough approximation.

The factor a may, of course, also model the effect of

different perceptions of waiting time and in-vehicle time. Since the link probabilities (13) are independent of the units in which f_l is specified, it is always possible to scale the frequencies by a factor 1/a. Without loss of generality we can therefore assume a=1 in the following sections.

In order to take advantage of the special case (12) and (13) for the solution of our initial problem P1, we need to express the strategy $\bar{L}$ in terms of the 0-1 variables x_l

$$x_l = \begin{cases} 0 & ,\text{if } l \notin \bar{L} \\ & \\ 1 & ,\text{if } l \in \bar{L} \end{cases} \qquad ,l \in L. \tag{14}$$

We can now rewrite P1 using (12), (13) and (14)

$$\text{Min} \quad \sum_{l \in L} t_l v_l + \sum_{i \in I} \frac{v_i}{\sum_{l \in L_i^+} f_l x_l} \tag{P4}$$

subject to

$$v_l = \frac{x_l f_l}{\sum_{l' \in L_i^+} f_{l'} x_{l'}} v_i \qquad ,l \in L_i^+, i \in I, \tag{15}$$

$$v_i = \sum_{l \in L_i^-} v_l + g_i \qquad ,i \in I, \tag{16}$$

$$v_i \geq 0 \qquad ,i \in I, \tag{17}$$

$$x_l = 0 \text{ or } 1 \qquad ,l \in L. \tag{18}$$

If we compare problem P4 to P2 and P3, we see that we have drastically reduced the size of the problem. The number of variables and constraints are only of the order of the number of links and nodes of the network and no longer imply an explicit enumeration of all possible link subsets. On the other hand, the complexity of the problem seems to have increased considerably. Problem P4, at first sight, consists of a non-linear objective function subject to non-linear constraints, and the variables are partly continuous, partly integer. However, we shall demonstrate that this problem may be reduced to a much simpler linear programming problem by observing the following.

First, the non-negativity constraints for the node volumes (17) may be replaced by non-negativity constraints for the link volumes because of (15)

$$v_l \geq 0 \qquad ,l \in L. \tag{19}$$

Next, by summing the constraints (15) for $l \in L_i^+$ we obtain

$$\sum_{l \in L_i^+} v_l = v_i \qquad ,i \in I. \tag{20}$$

which may be used to substitute v_i in (16) to yield conservation of flow constraints (3).

Then, by introducing new variables w_i that denote the

total waiting time for all trips at node i

$$w_i = \frac{v_i}{\sum_{l \in L_i^+} f_l x_l} \qquad ,i \in I, \tag{21}$$

the variables v_i, $i \in I$, may be substituted.

The arguments leading to the justification of (19), (20) and (21) demonstrate the following.

Proposition 3: Problem P4 is equivalent to problem P5

$$\text{Min} \quad \sum_{l \in L} t_l v_l + \sum_{i \in I} w_i \tag{P5}$$

subject to (3), (18), (19) and

$$v_l = x_l f_l w_i \qquad ,l \in L_i^+, i \in I. \tag{22}$$

As we can see, the objective function is now linear and the node volumes v_i, $i \in I$, no longer appear explicitly in the formulation. The 0-1 variables x_l are only used in constraints (22), which, at the same time, are the only non-linear constraints of problem P5. These constraints may be relaxed by replacing (22) with

$$v_l \leq f_l w_i \qquad ,l \in L_i^+, i \in I, \tag{23}$$

to yield the linear programming problem

$$\text{Min} \quad \sum_{l \in L} t_l v_l + \sum_{i \in I} w_i \tag{P6}$$

subject to (3), (19) and (23).

Proposition 4: Problem P6 is equivalent to problem P5.

Proof: Every extreme point of problem P6 has at most as many inequality constraints that are strictly satisfied as there are links in the network. Since for every link there are exactly two inequalities, (19) and (23), that may not be satisfied with equality at the same time (mutually exclusive), there is at least one strict inequality associated with each link. We conclude that for every link either (19) or (23) is satisfied with equality. Hence, every extreme point of P6 is a feasible point of P5. Since P6 is a relaxation of P5, the two problems must be equivalent. □

Having finally cast the model as a linear programming problem, we can now examine its dual

$$\text{Max} \quad \sum_{i \in I} g_i u_i \tag{P7}$$

subject to

$$u_j + t_l + m_l \geq u_i \qquad ,l=(i,j) \in L, \tag{24}$$

$$\sum_{l \in L_i^+} f_l m_l = 1 \qquad ,i \in I, \tag{25}$$

$$m_l \geq 0 \qquad ,l \in L. \tag{26}$$

Let (v^*,w^*) and (u^*,m^*) denote any optimal solutions of problems P6 and P7 which always satisfy the complementary slackness conditions

$$(v_l^* - f_l w_i^*) m_l^* = 0 \qquad ,l \in L_i^+, i \in I, \tag{28}$$

and

$$(u_j^* + t_l + m_l^* - u_i^*) v_l^* = 0 \qquad ,l=(i,j) \in L. \tag{29}$$

So far we have only considered the case in which all links of the network involve waiting, that is, they have a finite frequency f_l. The case in which a subset of the links $\underline{L} \subseteq L$ involves no waiting may be handled in two ways. A simple approach is to set $f_l = M$, where M is a sufficiently large number, so that the resulting waiting time 1/M can be ignored. It is also possible to modify problems P4, P5 and P6, similarly to allow for links without waiting component. In problems P4 and P5 this results in dropping the constraints (22) and (23) for $l \in \underline{L}$ and adding explicitly non-negativity constraints for w_i, $i \in I$. In the dual problem P6, constraint (24) applies only for links $l \notin \underline{L}$ and is replaced by $u_j + t_l \geq u_i$ for $l \in \underline{L}$. Constraint (25) becomes

$$\sum_{l \in \underline{L}_i^+} f_l m_l \leq 1 \qquad ,i \in I.$$

The complementary slackness conditions (28) and (29) reduce to $(u_j^* + t_l - u_i^*) v_l^* = 0$ for $l \in \underline{L}$.

Finally, it is worth while noting that, both in its primal and dual formulations, the transit assignment model presented here has a close resemblance to the shortest path assignment problem. In fact, the latter corresponds to a special case of the transit assignment problem for which

none of the links of the network involves waiting, that is, $\underline{L} = L$. It is therefore not surprising that, as we shall see next, there is also a resemblance between the solution algorithms of the two problems.

5. The Algorithm

In this section we present an algorithm that solves the transit assignment problem P4. The algorithm is composed of two parts. In a first pass, from the destination node to all origins, the optimal strategy $\bar{L}^*$ and the expected total travel times u_i^* from each node $i \in I$ to the destination node r are computed. In a second pass, from all origins to the destination, the demand is assigned to the network according to the optimal strategy.

The algorithm is stated below:

Part 1: Find Optimal Strategy

1.1 (Initialization) $u_i := \infty$, $i \in I-\{r\}$; $u_r := 0$;
$f_i := 0$, $i \in I$;
$S := L$; $\bar{L} := \emptyset$.

1.2 (Get next link) If $S = \emptyset$ then STOP,
else find $l=(i,j) \in S$ which satisfies
$u_j + t_l \leq u_{j'} + t_{l'}$, $l' \in S$;
$S := S - \{l\}$.

1.3 (Update node label) If $u_i \geq u_j + t_l$ then

$$u_i := \frac{f_i u_i + f_l (u_j + t_l)}{f_i + f_l},$$

$f_i := f_i + f_l$, $\bar{L} := \bar{L} + \{l\}$;
go to step 1.2.

Part 2: Assign Demand According to Optimal Strategy

2.1 (Initialization) $v_i := g_i$, $i \in I$;

2.2 (Loading) Do for every link $l \in L$, in decreasing order of $(u_j + t_l)$:

$$\text{if } l \in \bar{L} \text{ then } v_l := \frac{f_l}{f_i} v_i,$$

$$v_j := v_j + v_l,$$

$$\text{else } v_l := 0.$$

In step 1.1 the node labels u_i, i.e. the expected travel times to reach the destination, are set to infinity for all nodes except the destination, for which it is initialized to zero. The auxiliary variables f_i, $i \in I$, that contain the combined frequencies of all selected links at node i, are initialized to zero. The set S is used to identify the links that have not yet been examined, and the set $\bar{L}$ is used to identify the optimal strategy.

In step 1.2 the link nearest to the destination is selected among the links not yet examined. The time considered is $u_j + t_l$, that is, the time from the i-node of the link to the destination not including the waiting time at node i. If this time is smaller than the current time associated with node i, u_i, link l is included in the optimal strategy and u_i and f_i are updated according to the formulas given in step 1.3. Note that when the label of a node i, u_i, is improved for the first time, we have $f_i u_i = 0 \cdot \infty$. In

order to keep the algorithm as compact as possible we adapt the convention $0 \cdot \infty = 1$ whenever this case occurs. (When $a \neq 1$ is used in (12), this convention becomes $0 \cdot \infty = a$). Part 1 of the algorithm terminates when all links have been examined.

In the second part of the algorithm the demand from node i to the destination, g_i, is assigned to the network according to the strategy $\bar{L}$. This is done by assigning to each link $l \in \bar{L}$ the proportion of the node volume v_i that corresponds to its frequency. Since the links are processed in reverse topological order (decreasing $u_j + t_l$), the node volumes may be updated in parallel, hence making it possible to examine every link only once.

In order to prove that the solution $(\hat{u}, \hat{v})$ obtained by this algorithm is an optimal solution of problem P4, we proceed to show that $\hat{u}$ is a feasible solution of the dual problem P6, $\hat{v}$ is a feasible solution of the primal problem P5 and that $(\hat{u}, \hat{v})$ satisfy the complementary slackness conditions for problems P5 and P6. Optimality then results directly from the theorem of complementary slackness in linear programming.

Proposition 5: $\hat{u}$ **is a feasible solution of (dual) problem P6.**

Proof: Let u_i^l and u_j^l be the values of u_i and u_j after step 1.3 has been performed for link $l=(i,j)$. Then

$$u_i^l > u_j^l + t_l \quad \text{,if } l \in \bar{L},$$
$$u_i^l \leq u_j^l + t_l \quad \text{,else.} \tag{30}$$

As the links are examined in increasing order of u_j+t_l and $t_l \geq 0$, it follows that u_j will not be modified any more, therefore $u_j^l=\hat{u}_j$, $l \in L$. This is not true for u_i which may still be modified in step 1.3. Since u_i can only decrease, we have $\hat{u}_i \leq u_i^l$, $l \in L$. On the other hand, $\hat{u}_i=u_i^{l'}$, where $l' \in \bar{L}_i^+$ is the link included last into $\bar{L}$, which leads to

$$\hat{u}_i > u_j^{l'} + t_l, \geq u_j^l + t_l \quad \text{,} l \in \bar{L}_i^+. \tag{31}$$

Thus, it follows that $\hat{u}$ satisfies

$$\hat{u}_i > \hat{u}_j + t_l \quad \text{,if } l \in \bar{L},$$
$$\hat{u}_i \leq \hat{u}_j + t_l \quad \text{,else.} \tag{32}$$

Furthermore, since we have shown that u_j will not change after it has been used in step 1.3, $\hat{u}_i$ may be expressed in terms of $\hat{u}_j$

$$\hat{u}_i = \frac{\sum_{l \in \bar{L}_i^+} f_l(\hat{u}_j+t_l) + 1}{\sum_{l \in \bar{L}_i^+} f_l} \quad \text{,} i \in I. \tag{33}$$

In (33) we easily recognize the first term as the weighted average of the times for each link and the second term as the waiting time at node i. The dual variable $\hat{m}$ may be constructed as

$$\hat{m}_l = \begin{cases} \hat{u}_i - \hat{u}_j - t_l & \text{,} l \in \bar{L}, \\ 0 & \text{,else.} \end{cases} \tag{34}$$

We can rewrite (33) as

$$\sum_{l \in \bar{L}_i^+} f_l(\hat{u}_i - \hat{u}_j - t_l) = 1 \qquad ,i \in I. \tag{35}$$

Now, by inserting (34) into (35), it follows that

$$\sum_{l \in L_i^+} f_l \hat{m}_l = 1 \qquad ,i \in I. \tag{36}$$

Since by construction $(\hat{u},\hat{m})$ satisfies (24) and (26), it follows that this is a feasible solution of problem P6. □

Proposition 6: $\hat{v}$ is a feasible solution of (primal) problem P5.

Proof: Let $\hat{f}_i$, $i \in I$, be the combined frequencies obtained by part 1 of the algorithm. According to step 1.3 they satisfy

$$\hat{f}_i = \sum_{l \in \bar{L}_i^+} f_l \qquad ,i \in I. \tag{37}$$

Let the variables $\hat{w}_i$, $i \in I$, be defined as

$$\hat{w}_i = \frac{\hat{v}_i}{\hat{f}_i} \qquad ,i \in I. \tag{38}$$

We can now express $\hat{v}_l$ in terms of $\hat{w}_i$ by using (38) and step 2.2 of the algorithm

$$\hat{v}_l = \begin{cases} f_l \hat{w}_i & ,\text{if } l \in \bar{L}, \\ 0 & ,\text{else.} \end{cases} \tag{39}$$

Since the link volumes satisfy the conservation of flow

equations and are non-negative by construction, it follows that $(\hat{v},\hat{w})$ is a feasible solution of problem P5. □

Proposition 7: $(\hat{u},\hat{v})$ are optimal.

Proof: Links included in strategy $\bar{L}$ satisfy

$$\hat{v}_l - f_l \hat{w}_i = 0 \qquad ,l=(i,j)\in\bar{L}, \tag{40}$$

and

$$\hat{u}_j + t_l + \hat{m}_l - \hat{u}_i = 0 \qquad ,l=(i,j)\in\bar{L}. \tag{41}$$

where for links not included

$$\hat{v}_l = 0, \quad \hat{m}_l = 0 \qquad ,l\notin\bar{L}. \tag{42}$$

Hence the solutions $(\hat{v},\hat{w})$ and $(\hat{u},\hat{m})$ satisfy the complementary slackness conditions (28) and (29) and are therefore optimal solutions of problems P5 and P6. Since $(\hat{v},\hat{w})$ is an extreme point solution of P5, $(\hat{v},\bar{L})$ is an optimal solution of problem P4 by proposition 3. □

The case, in which a subset of links, $\underline{L} \subseteq L$ does not involve waiting for service, may again be handled by either setting the frequencies f_l of these links to a very large number, or by adapting the algorithm to this special case. The latter approach yields two slight modifications of steps 1.3 and 2.2 of the algorithm. If $l\in\underline{L}$, the updating of the variables in step 1.3 becomes

$$u_i := u_j + t_l \ ,$$
$$f_i := \infty \ , \quad \bar{L}_i^+ := \{l\}.$$

If a link $l\in\underline{L}$ is part of the optimal strategy, its volume is

calculated as $v_1 := v_i$ in step 2.2. No proof is given for this modified version of the algorithm, since it follows the same lines as the proofs for propositions 5, 6 and 7.

The following example shows how the algorithm is applied to the small network used as an example in sections 2 and 3. Since the waiting times and the link probabilities given in Table 1 correspond to the special case (12) and (13), it is possible to assign frequencies f_1 to the links. Figure 7 shows the network together with the link times and frequencies (t_1, f_1).

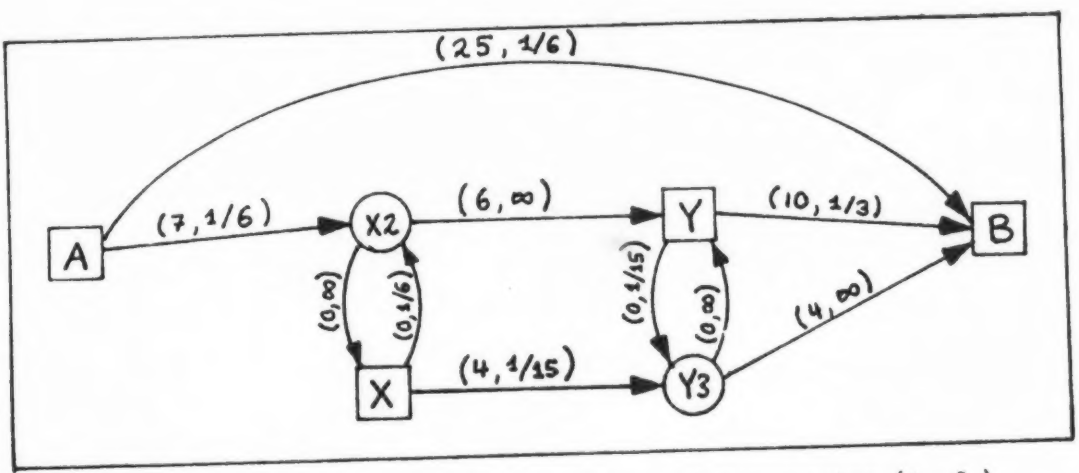

Figure 7: Transit network with link times and frequencies (t_1, f_1).

Note that some links have no waiting component ($f_1 = \infty$), the modified version of the algorithm is therefore used. Table 2 shows how the optimal strategy is computed (part 1 of the algorithm).

#	Node Labels (u_i, f_i) A	X2	X	Y3	Y	B	Link with Min u +t (i,j)	f_i	u_j+t_a	$\in \bar{L}$
1	∞,0	∞,0	∞,0	∞,0	∞,0	0,0	(Y3,B)	∞	4.0	yes
2	"	"	"	4.0,∞	"	"	(Y,Y3)	1/15	4.0	yes
3	"	"	"	"	19.0,1/15	"	(X,Y3)	1/15	8.0	yes
4	"	"	23.0,1/15	"	"	"	(Y,B)	1/3	10.0	yes
5	"	"	"	"	11.5,2/5	"	(Y,Y3)	∞	11.5	no
6	"	"	"	"	"	"	(X2,Y)	∞	17.5	yes
7	"	17.5,∞	"	"	"	"	(X,X2)	1/6	17.5	yes
8	"	"	19.0,7/30	"	"	"	(X2,X)	∞	19.0	no
9	"	"	"	"	"	"	(A,X2)	1/6	24.5	yes
10	30.5,1/6	"	"	"	"	"	(A,B)	1/6	25.0	yes
	27.75,1/3	17.5,∞	19.0,7/30	4.0,∞	11.5,2/5	0,0				

Table 2: Find Optimal Strategy for Example Network.

In Table 3 the demand, one trip from A to B, is assigned to the network (part 2 of the algorithm).

#	Link (i,j)	Volume	Node Volumes A	X2	X	Y3	Y	B
			1.00	0.00	0.00	0.00	0.00	-1.00
10	(A,B)	0.50	1.00	0.00	0.00	0.00	0.00	-0.50
9	(A,X2)	0.50	1.00	0.50	0.00	0.00	0.00	-0.50
8	(X2,X)	0.00	1.00	0.50	0.00	0.00	0.00	-0.50
7	(X,X2)	0.00	1.00	0.50	0.00	0.00	0.00	-0.50
6	(X2,Y)	0.50	1.00	0.50	0.00	0.00	0.50	-0.50
5	(Y,Y3)	0.09	1.00	0.50	0.00	0.00	0.50	-0.50
4	(Y,B)	0.42	1.00	0.50	0.00	0.00	0.50	-0.08
3	(X,Y3)	0.00	1.00	0.50	0.00	0.00	0.50	-0.08
2	(Y3,Y)	0.08	1.00	0.50	0.00	0.08	0.50	-0.08
1	(Y3,B)	0.08	1.00	0.50	0.00	0.08	0.50	0.00

Table 3: Assign Demand on Example Network.

The solution found by the algorithm corresponds to the optimal strategy displayed in Figure 2.

We examine next the complexity of the algorithm. Let n and m be the number of nodes and links in the network considered. The following table shows the complexity of every step of the algorithm and the number of times it is executed.

Step	Complexity	#Times
1.1	O(n+m)	1
1.2	O(m)	m
1.3	O(1)	m
2.1	O(n)	1
2.2	O(m)	1

Table 4. Complexity of the Steps of the Algorithm

The complexity of the entire algorithm is determined by step 1.2 and is $O(m^2)$. Note that in step 2.2 no additional computations are needed to establish the order in which the links are processed, since it is the inverse of the order used in part 1 of the algorithm. Step 1.2 can, of course, be implemented by using a heap structure, thus reducing its complexity to O(log m) and the complexity of the algorithm to O(m log m).

If the network contains more than one destination node, the algorithm is applied once for each destination. In this case working memory space can be saved by modifying the algorithm slightly so that the variables v_l, leL, are used directly to store the accumulated link volumes for all destinations as follows:

(I) Initialize link volumes $v_l := 0$, $l \in L$.

(II) Assign the demand to each destination using the algorithm above with step 2.2 modified in the following way:

Do for every link $l \in L$, in decreasing order of $(u_j + t_l)$:

$$\text{if } l \in \bar{L} \text{ then } v_l := v_l + \frac{f_l}{f_i} v_i,$$

$$v_j := v_j + \frac{f_l}{f_i} v_i.$$

6. Model Extensions

We now turn our attention to the case of the transit assignment problem in which the link travel times t_l are no longer constants, but are continuous, increasing functions $t_l(v_l)$ of the total volume on link l. This functional relation between link travel time and link volume, $t_l(v_l)$, may be used to model the transit vehicle's performance under different load conditions, such as the slowing down of a fully loaded bus on an uphill road. In a somewhat limited way, it may also be used to model congestion or crowding effects within the transit vehicles. In this case, $t_l(v_l)$ would not only be the travel time, but it would also include a discomfort effect, expressed in units of generalized time or cost, that increases with the number of passengers aboard. This approach allows us to model some of the consequences of transit vehicle capacity by using discomfort functions similar to the volume delay functions used in the equilibrium road assignment to model street capacities. However, there are some limitations to this approach that have to be kept in mind:

- In this model, all passengers aboard a transit vehicle suffer the same degree of discomfort. In reality the passenger that boards the vehicle first has a better chance of getting a seat and experiences therefore a different level of discomfort than a passenger that boards the vehicle when no more seating space is available.

- The waiting time is not directly affected by the transit volumes. Every passenger may board the first arriving vehicle and queuing delays at stops because of full vehicles are not considered by the model.

- The time the vehicle spends at the transit stop may depend in reality on the number of passengers that alight and board at that stop. This dependence is not considered either by this model.

Let v_l^r denote the volume on link $l \in L$ that is destined to destination node $r \in R$. Since the total volume v_l on a link is the sum of the volumes v_l^r to all different destinations $r \in R$,

$$v_l = \sum_{r \in R} v_l^r \qquad , l \in L, \tag{43}$$

it is no longer possible to separate the problem by destination.

Let the demand from node $i \in I$ to destination $r \in R$ be denoted g_i^r. Let also K_r denote the set of all feasible strategies to reach destination r. h_i^k is the part of the demand g_i^r that is assigned to the network using strategy $k \in K_r$. Obviously we have conservation of flow

$$\sum_{k \in K_r} h_i^k = g_i^r \qquad , i \in I,\ k \in K_r,\ r \in R. \tag{44}$$

We assume, in the same way we did for constant link times, that every traveler chooses the strategy that minim-

izes the expected travel time for his trip. But now the travel times depend on the total link volumes and therefore on the strategies chosen by the other travelers. Mathematically, the travelers' behavior can be characterized by equilibrium conditions similar to the ones derived from Wardrop's second principle which are used for the car traffic equilibrium assignment [18,22]. Let s_i^k denote the expected travel time from node i to destination r using strategy $k \in K_r$ and let u_i^r denote the minimum expected travel time from i to r; the equilibrium conditions may now be stated as

$$\left.\begin{array}{l} s_i^{k*} = u_i^{r*}, \text{ if } h_i^{k*} > 0 \\ s_i^{k*} \geq u_i^{r*}, \text{ if } h_i^{k*} = 0 \end{array}\right\} \quad ,i \in I,\ k \in K_r,\ r \in R. \tag{45}$$

Proposition 8: Any optimal solution of problem P8 is a flow that satisfies the equilibrium conditions (45):

$$\text{Min} \sum_{l \in L} \int_0^{v_l} t_l(x)dx + \sum_{i \in I} \sum_{r \in R} w_i^r \tag{P8}$$

subject to (43) and

$$v_l^r \leq f_l w_i^r \quad ,l \in L_i^+,\ i \in I,\ r \in R, \tag{46}$$

$$\sum_{l \in L_i^+} v_l^r = \sum_{l \in L_i^-} v_l^r + g_i^r \quad ,i \in I,\ r \in R, \tag{47}$$

$$v_l^r \geq 0 \quad ,l \in L,\ r \in R. \tag{48}$$

Proof: We first note that the constraints of problem P8 are separable by destination. The extreme points of the feasible set associated with a destination r correspond to the set of feasible strategies to reach destination r, as we have seen in proposition 4. It is therefore possible to restate problem P8 in terms of the volumes v_l^k associated with each feasible strategy $k \in K_r$.

$$\text{Min} \sum_{l \in L} \int_0^{v_l} t_l(x)dx + \sum_{r \in R} \sum_{k \in K_r} \sum_{i \in I} w_i^k \qquad \text{(P9)}$$

subject to (44) and

$$v_l^k = f_l x_l w_i^k \qquad , l \in L_i^+,\ i \in I,\ k \in K_r,\ r \in R, \qquad (49)$$

$$\sum_{l \in L_i^+} v_l^k = \sum_{l \in L_i^-} v_l^k + h_i^k \qquad , i \in I,\ k \in K_r,\ r \in R, \qquad (50)$$

$$v_l = \sum_{i \in I} \sum_{k \in K_r} v_l^k \qquad , l \in L, \qquad (51)$$

$$w_i^k \geq 0\ ,\ h_i^k \geq 0 \qquad , i \in I,\ k \in K_r,\ r \in R. \qquad (52)$$

The constants x_l^k are 1 if link l is part of strategy k, 0 otherwise. Let $\hat{m}_l^k$, $\hat{s}_i^k$ and $\hat{u}_i^r$ denote the dual variables that correspond to the constraints (49), (50) and (44). The Kuhn - Tucker conditions for an optimal solution (v^*, w^*, h^*) of problem P9 are then derived as

$$\hat{s}_j^{k^*} + t_l(v_l^*) + \hat{m}_l^{k^*} \begin{cases} = \hat{s}_i^{k^*}, & \text{if } h_i^{k^*} > 0 \\ \geq \hat{s}_i^{k^*}, & \text{if } h_i^{k^*} = 0 \end{cases} \quad l=(i,j) \in L, k \in K_r, r \in R, \qquad (53)$$

$$\sum_{l \in L_i^+} \hat{m}_l^{k^*} x_l^k f_l = 1 \qquad , i \in I,\ k \in K_r,\ r \in R \tag{54}$$

and

$$\left.\begin{aligned} \hat{s}_i^{k^*} &= \hat{u}_i^{r^*} && , \text{if } h_i^{k^*} > 0 \\ \hat{s}_i^{k^*} &\geq \hat{u}_i^{r^*} && , \text{if } h_i^{k^*} = 0 \end{aligned}\right\} \quad , i \in I,\ k \in K_r,\ r \in R. \tag{55}$$

In (53) and (54) we recognize the same structure as we have seen in (24) and (25). The variables $\hat{s}_i^{k^*}$ correspond therefore to the expected travel time from node i to destination r using strategy k; that is,

$$\hat{s}_i^{k^*} = s_i^{k^*} \qquad , i \in I,\ k \in K_r,\ r \in R. \tag{56}$$

Since (55) implies $\hat{u}_i^{r^*} = \min_{k \in K_r} \hat{s}_i^{k^*}$, we also have

$$\hat{u}_i^{r^*} = u_i^{r^*} \qquad , i \in I,\ r \in R \tag{57}$$

and hence (55) is equivalent to (45). □

Since problem P8 is a convex minimization problem, it may be solved by using any of the standard techniques available for this class of problems. In the following, we outline briefly a solution algorithm that is based on the Frank and Wolfe method [11]. Since it does not require the explicit storage of the destination dependent volumes v_l^r, it is especially suited to handle large transit networks with many destination nodes. We note that the linear approximation of

the objective function of problem P8 results in a problem that is separable by destination. For each destination the resulting subproblem is equivalent to the problem with constant link times as discussed in section 4, and may therefore be solved efficiently using the algorithm given in section 5. The resulting algorithm to solve problem P8 may be outlined as follows:

0. (Initialization) Find a feasible solution (v^0, w^0).
Set $k := 0$.

1. (Subproblem) Set $k := k + 1$.
Compute $(\hat{v}, \hat{w})$ by applying the algorithm of section 5 with link times $t_l = t_l(v_l^{k-1})$ for each destination $r \in R$.

2. (Line Search) Find λ^k which minimizes the objective function of P8 on the line segment $(1-\lambda)(v^{k-1}, w^{k-1}) + \lambda(\hat{v}, \hat{w})$, $0 \leq \lambda \leq 1$.

3. (Update) Set $(v^k, w^k) := (1-\lambda^k)(v^{k-1}, w^{k-1}) + \lambda^k(\hat{v}, \hat{w})$.
If $||(v^k, w^k) - (v^{k-1}, w^{k-1})|| < \epsilon$ then STOP
else go to step 1.

7. Conclusions

In this paper we have developed a new approach to the fixed route transit assignment problem. The main difference between the proposed model and existing approaches to the transit assignment is twofold. First and most important, in our model the traveler does not choose a path, as in all existing methods, but more realistically he selects a strategy. While in all existing methods the transit traveler's route choice is limited to one path, be it defined as a sequence of route segments, links or transfer points, the strategies considered in our approach allow the transit rider to select any subset of route segments that lead to his destination and let the vehicle that arrives first determine which of the alternative routes is actually taken on an individual trip. This concept of optimal strategies allows modelling of the traveler's behavior much more realisticly. Second, the transit assignment problem has been formulated for the first time as a mathematical programming problem that is based on sound behavioral assumptions and may be solved without recourse to heuristic approaches. In its general form, the model allows for arbitrary vehicle interarrival time distributions. Using the special case of combining frequencies, the problem size reduces drastically and results in a linear programming formulation that has similarities with the shortest path assignment problem. The proposed solution algorithm solves the problem in polynomial time. Since time and memory requirements of the algorithm

are comparable to those of the shortest path assignment, the algorithm may be applied even to very large transit networks.

We have shown that the model is not restricted to constant link times. It may be generalized to include link time functions $t_l(v_l)$ that depend on the total volume on the links. In this case the solution may be characterized by equilibrium conditions similar to those derived from Wardrop's second principle. The existence of an equivalent convex optimization formulation was demonstrated and an efficient solution algorithm that is based on the Frank and Wolfe method was outlined.

Looking at the simplicity of the proposed models and algorithms, it is surprising that this model formulation was not developed earlier. Several reasons may have contributed to this fact:

- Shortest path algorithms have been well known for a long time. All road assignment methods are based on path choice and use shortest paths in one form or another. It is therefore not surprising that, when the transit assignment problem was considered, it was taken for granted that the transit rider's route choice is also based on shortest paths.

- Historically, shortest path algorithms and assignment methods always worked with one origin node at a time,

computing results for all destinations simultaneously. Since the shortest path problem and road assignment problems are symmetric with respect to origin and destination, this choice is purely arbitrary; that is, one could as well compute shortest paths from all origins to one destination at the same time. In the case of the transit assignment problem this symmetry no longer holds due to the waiting phenomenon; the optimal strategy can only be efficiently computed from all origins to one destination at a time. Unfortunately, this is not the way we are accustomed to.

The algorithm described in this paper has been implemented as part of the EMME/2 system, an interactive - graphic planning tool for urban transportation planning, developed at the Centre de Recherche sur les Transports de l'Université de Montréal [2]. The data structure has been adapted to avoid the explicit construction of the links and nodes of the generalized transit network. Table 5 shows the execution times for the transit assignment with constant link times for the networks of the cities of Winnipeg, Canada, and Trujillo, Peru. The program is written in FORTRAN and is installed on two computers: the CDC Cyber 173 of the University of Montreal and on a Pixel 100/AP 16-bit microcomputer. The execution times are only slightly higher than the times needed for a shortest path assignment for a network of the same size.

City:	Winnipeg	Trujillo
No. of transit routes:	67	33
No. of route segments:	4205	2902
No. of origin/destination zones:	147	67
No. of generalized links:	15684	10778
Execution times (CPU seconds):		
CDC Cyber 173:	171.78	60.32
Pixel 100/AP:	1179.20	394.70

Table 5. Execution Times for Transit Assignment Algorithm

Since our model includes as a special case the road equilibrium assignment model, it may be used to unify the formulation of transit and road assignment models. Of course it may also be applied directly to other assignment problems in which waiting for service has to be considered, such as moving freight on a truck or railway network.

This model may serve as a starting point for future research. Several extensions of the model are possible and have a high potential for providing new insight into the area of public transportation modeling.

In section 6 we mentioned several limitations of the non-linear model when modeling congestion effects. Most of these difficulties could be overcome by having the link time t_l not only depend on the volume of link l, $t_l(v_l)$, but also

of the volumes of other links $t_l(v)$. This model may be stated in the form of a variational inequality. Unfortunately, a unique solution to the problem is only ensured if the mapping $t(v)$ is strictly monotone. Since the phenomena we try to model lead to non-monotone mappings $t(v)$, multiple equilibria may occur, rendering the problem very difficult to solve. Another interesting approach to the modeling of congestion effects would consist of letting the frequencies f_l be decreasing functions of the total volume aboard the vehicle $f_l(v)$. This "effective frequency" would take into account increased waiting times due to overcrowded vehicles that prevent boarding and force the passengers to wait for a subsequent vehicle. This extension of the model is not well understood yet.

Several authors have proposed transit network design models that are based on the optimization of the frequencies [8,15,20]. Replacing the simplistic route choice models that are used in these approaches by our model could lead to more realistic transit network design models.

Appendix A. Combined Waiting Time and Link Probabilities

Let $h_l(x)$ be the density function of the distribution of interarrival times (headways) of the vehicles on link l. Let $H_l(x)$ denote its cumulative

$$H_l(x) = \int_0^x h_l(t)dt \qquad ,l \in L. \tag{A1}$$

Assuming uniform arrival of passengers at the transit stop, the waiting time distribution for service on link l is

$$g_l(x) = 1 - H_l(x) \qquad ,l \in L, \tag{A2}$$

with the cumulative

$$G_l(x) = \int_0^x (1-H_l(t))dt \qquad ,l \in L. \tag{A3}$$

The combined waiting time for the first vehicle among links $l \in \bar{L}_i^+$ may be calculated as

$$W(\bar{L}_i^+) = \int_0^{\infty} \prod_{l \in \bar{L}_i^+} [1-G_l(x)] \; dx \tag{A4}$$

and the probability that link l is served first among $\bar{L}_i^+$ is

$$p_l(\bar{L}_i^+) = \int_0^{\infty} g_l(x) \prod_{\substack{l' \in \bar{L}_i^+ \\ l' \neq l}} [1-G_{l'}(x)]dx. \tag{A5}$$

References

1. Andreasson I., "A Method for the Analysis of Transit Networks", 2nd European Congress on Operations Research, edited by Mark Roubens, North Holland, 1976.

2. Babin A., Florian M., James-Lefebvre L. and Spiess H., "EMME/2: Interactive Graphic Method for Road and Transit Planning", Transportation Research Record 866 pp 1-9, 1982.

3. Chapleau R., "Rèseaux de Transport en Commun: Structure Informatique et Affectation", Ph.D. thesis, Dèp. I.R.O., Universitè de Montrèal, 1974.

4. Chriqui C., "Rèseau de Transport en Commun: Les Problémes de Cheminement et d'Accés", Ph.D. thesis, Dèp. I.R.O., Universitè de Montrèal, 1974.

5. Chriqui C. and Robillard P., "Common Bus Lines", Transportation Science 9, pp 115-121, 1975.

6. Dial R.B., "Transit Pathfinder Algorithm", Highway Research Record 205, pp 67-85, 1967.

7. Dial R.B., "A Probabilistic Multipath Traffic Assignment Model which Obviates Path Enumeration", Transportation Research Vol 5, 1971.

8. Fearnside K. and Draper D.P., "Public Transport Assignment - A New Approach", Traffic Engineering and Control, pp 298-299, November 1971.

9. Florian M., "A Traffic Equilibrium Model of Traffic by Car and Public Transit Modes", Transportation Research 10, pp 166-179, 1977.

10. Florian M. and Spiess H., "On Two Mode Choice / Assignment Models", Publication No. 177, Centre de Recherche sur les Transports, Universitè de Montrèal, 1981.

11. Frank M. and Wolfe P., "An Algorithm for Quadratic Programming", Naval Research Logistics Quarterly 3, pp 95-110, 1956.

12. General Motors Transportation System Center, "Interactive Graphic Transit Design System (IGTDS): User's Manual", General Motors Technical Center, Warren, MI, 1978.

13. Hasselström D., "Public Transportation Planning - A Mathematical Programming Approach", Ph.D. thesis, Department of Business Administration, University of Gothenburg, Sweden, 1981.

14. Lampkin W. and Saalmans P.D., "The Design of Routes, Service Frequencies and Schedules for a Municipal Bus Undertaking: A Case Study", Operations Research Quarterly, Vol 18, No 4, pp 375-397, 1967.

15. Last A. and Leak S.E., "Transept: a Bus Model", Traffic Engineering and Control, pp 14-20, January 1976.

16. Le Clercq F., "A Public Transport Assignment Method", Traffic Engineering and Control", pp 91-96, June 1972.

17. Mandl C.E., "Evaluation and Optimization of Urban Public Transportation Networks", presented at the 3rd European Congress on Operations Research, Amsterdam, Netherlands, 1979.

18. Nguyen S., "An Algorithm for the Traffic Assignment Problem", Transportation Science 8, pp 203-216, 1974.

19. Rapp M.H., Mattenberger P., Piguet S. and Robert-Grandpierre A., "Interactive Graphic System for Transit Route Optimization", Transportation Research Record 619, 1976.

20. Schèele C.E., "A Mathematical Programming Algorithm for Optimal Bus Frequencies", Institute of Technology, University of Linköping, Sweden, 1977.

21. UMTA/FHWA, "UTPS Reference Manual", U.S. Department of Transportation, 1977.

22. Wardrop J.G., "Some Theoretical Aspects of Road Traffic Research", Proc. Inst. Civil Engineers, Part II 1, pp 325-378, 1952.

Acknowledgements

The author benefitted from discussions with Michael Florian on the contents of this paper. Stefano Pallottino shared with the author his interest in the modelling of transit route choice and Linda James-Lefebvre helped with the style of the final presentation.